

Bayesian Analysis of Nested Logit Model by Markov Chain Monte Carlo*

Kajal Lahiri and Jian Gao
Department of Economics
University at Albany-SUNY
Albany, NY 12222
e-mail: klahiri@albany.edu
Tel: (518) 442-4758

April 26, 2001

*An earlier version of this paper was presented at the regional meeting of the International Society for Bayesian Analysis (ISBA) at Laguna Beach, April 5-7, 2001, and at the North American Summer Meeting of the Econometric Society, University of Maryland in College Park, June 21-24, 2001. We thank David Brownstone, Gary Koop, Marc Nerlove, Dale Poirier, and Arnold Zellner for many helpful comments. However, the responsibility for any errors is solely ours.

Abstract: We develop a Markov Chain Monte Carlo (MCMC) algorithm for estimating nested logit models in a Bayesian framework. Appropriate “heating target” and reparameterization techniques are adopted for fast mixing. For illustrative purposes, we have implemented the algorithm on two real-life examples involving 3-level structures. The first example involves Social Security’s disability determination process, Lahiri et al. (1995). The second one is taken from Amemiya and Shimono’s (1989) model of labor supply behavior of the aged. We applied a combination of various convergence criteria to ensure that the chain has converged to its target distribution.

Key words: Discrete choice, Random utility maximization, MCMC, Mixing speed.

JEL classification: C11, C25, H55, I12, J14.

1. Introduction

The nested multinomial logit (NMNL) model has many advantages over the multinomial logit (MNL) model. The NMNL model overcomes the *Independence of Irrelevant Alternatives* (IIA) property of the MNL model, can explore the sequential decision making processes, and improves upon the sequential logit model by allowing for correlations among error terms across different levels, see McFadden (1977, 1981). Not surprisingly, the model has been found to be very useful in many different areas of economics and public policy analysis.¹ Despite the advantages and popularity of NMNL model, serious difficulties in maximizing its likelihood function remain, and are not widely recognized.

In the early years of its development, most studies used the limited information maximum likelihood (LIML) approach to estimate the model. However, LIML estimates are not efficient and obtaining the correct standard errors of LIML estimates under diverse tree structures is not a trivial matter. Although the full information maximum likelihood (FIML) estimation is preferable to LIML, there are very few studies that report reliable FIML estimates of the NMNL model, particularly when the number of levels exceeds 2. In order to obtain reasonable FIML estimates, many authors find it necessary to impose *ad hoc* restrictions on the parameters; for example, the coefficients of inclusive values in different levels are generally assumed to be the same (Brownstone and Small, 1989). Another empirical puzzle is that LIML and FIML estimates are often widely different, even though both are theoretically consistent (Berkovec and Rust 1985; Cameron 1985). We believe that the difficulty comes from an extremely irregular surface of the maximum likelihood function. The severity of the difficulty, of course, depends on the model and data structure at hand. In short, attaining the global maximum of NMNL likelihood function is not guaranteed by the available FIML algorithms, and we believe this is the main reason why applications of NMNL model have been limited to rather simple two-level models.

Following Poirier (1996), we adopt a Bayesian approach for the estimation of NMNL models. The Bayesian approach offers many advantages over the classical methods. Different prior specifications can give the investigators more options in model development (Geweke 1999). As Poirier (1996) has noted, the use of priors

¹See, for instance, Amemiya and Shimono(1989), Berkovec and Rust (1985), Cameron (1985), Dubin (1998), Falaris (1984, 1987), Guimaraes et al. (1998), Hausman, et al. (1995), Hoffman and Duncan (1988), McFadden (1976, 1978, 1980), Newbold (1997), Weiler (1989), Train (1980) and Train et al. (1989).

merely formalizes the already existing non-Bayesian practice of using non-data-based information in various informal ways. We will also argue that the Bayesian approach is a natural choice in the case of the NMNL model where maximum likelihood (ML) estimation can be notoriously unreliable.

Poirier (1996) developed a Bayesian framework for the NMNL model. However, analysis of the posterior distribution of the NMNL model is not straightforward. The Monte Carlo integration technique seems to work well for the MNL model (see Koop and Poirier, 1993), but finding an appropriate importance sampling density for the NMNL model is difficult (see Geweke, 1989). With the increasing availability of computing power and the development of Markov Chain Monte Carlo (MCMC) methods, the Bayesian approach to analyze such models has become feasible. The Metropolis and Hastings (M-H) algorithm, and its special case, the Gibbs Sampler, have been used in recent years to handle a wide variety of otherwise intractable statistical problems². The Gibbs sampler does not help in estimating the NMNL model because its conditional posterior distribution is intractable. In addition, the difficulty of analyzing the posterior or the likelihood function increases exponentially with respect to the number of levels of the tree structure. As a result, the M-H algorithm seems to be a logical choice in analyzing the NMNL model.

In this study, we develop a practical Markov Chain Monte Carlo (MCMC) method for the Bayesian analysis of NMNL models. We pay special attention to increase the mixing speed of the chain. For the sake of illustration, we implemented this simulation framework on two real life examples involving 3-level nested structures. Through an elaborate convergence diagnostic analyses, we ensure that the simulation results are reliable.

The paper is organized as follows: We will introduce the classical NMNL framework and the estimation issues in Section 2. Section 3 reviews the Bayesian framework. In Section 4 we will give a brief background for MCMC and discuss some of the implementation issues. In section 5 we present two empirical examples to illustrate the application of the simulation method. Finally, Section 6 summarizes the main results of the paper.

²See, for example, Casella and George (1992), Geyer (1992), Geyer and Thompson (1992, 1995), Besage and Green (1993), Smith and Roberts (1993), Tierney (1994), Chib and Greenberg (1996), Chib, Greenberg and Chen (1998), Geweke (1999), and the rich discussions accompanying many of these papers.

2. Classical NMNL Model

In order to overcome the IIA property of the MNL model and to allow for the sequential decision making process, McFadden (1977, 1978) generalized the MNL model based on random utility maximization by letting the so-called inclusive value parameters to have coefficients other than one, and proposed a new model known as the nested multinomial logit model (NMNL).

Suppose individual i (1, 2, ..., N) has j (1, 2, ..., M) choices, and the indirect utility function is expressed as

$$U_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad (2.1)$$

where U_{ij} is the utility attained by individual i who chooses choice j , and μ_{ij} is a function of all the measured characteristics. In practice, μ_{ij} is often a linear function of the individual and/or the choice specific characteristics, that is $\mu_{ij} = X_{ij}\beta_j$, and ε_{ij} is the residual that captures the effects of unmeasured variables, personal idiosyncracies, imperfections in perception and maximization and so on. Following McFadden (1981), we assume that the error terms have the following Type B extreme-value distribution (see also Amemiya 1985):

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M) = \exp \left\{ - \sum_{\omega \in D} c_\omega \left[\sum_{\kappa \in E_\omega} b_\kappa \left(\sum_{s \in F_\kappa} a_s \left(\sum_{j \in G_s} \exp(-\varepsilon_j / \rho_s) \right)^{\rho_s / \rho_\kappa} \right)^{\rho_\kappa / \rho_\omega} \right]^{\rho_\omega} \right\}, \quad (2.2)$$

where $\rho_s, \rho_\kappa, \rho_\omega$ are the coefficients of the inclusive values and satisfy $0 < \rho_s, \rho_\kappa, \rho_\omega < 1$; D, E_ω, F_κ and G_s are branch index set, and a_s, b_κ, c_ω are the branch indices,

$$a_s \text{ or } b_\kappa \text{ or } c_\omega = \begin{cases} 1 & \text{if } s \in F_\kappa \text{ or } \kappa \in E_\omega \text{ or } \omega \in D \text{ respectively,} \\ 0 & \text{otherwise.} \end{cases}$$

We suppress the index i (denote $\mu = (\mu_1, \mu_2, \dots, \mu_M)$, $\rho = (\rho_s, \rho_\kappa, \rho_\omega)$, and $\mu_j = X\beta'_j$) and define:

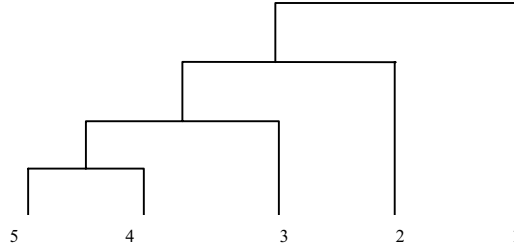
$$H(\kappa, \rho) = \sum_{\omega \in D} c_\omega \left[\sum_{\kappa \in E_\omega} b_\kappa \left(\sum_{s \in F_\kappa} a_s \left(\sum_{j \in G_s} \exp(\mu_j / \rho_s) \right)^{\rho_s / \rho_\kappa} \right)^{\rho_\kappa / \rho_\omega} \right]^{\rho_\omega}. \quad (2.3)$$

It is easy to verify that,

$$P_j(\mu, \rho) = \frac{\partial H / \partial \mu_j}{H(\mu_1, \mu_2, \dots, \mu_M)}, \quad j = 1, 2, \dots, M. \quad (2.4)$$

To appreciate the structure of the general NMNL model better, and in anticipation of the illustrative examples to be presented later, let us consider a 4-level NMNL model which is a special case of equation (2.2). The tree structure is depicted in Figure 1:

Figure 1. A 4-level Nested Structure



In this case $M = 5$, and the distribution function of error terms simplifies to

$$f(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5) = \exp\{-[\exp(-\varepsilon_1) + (\exp(-\varepsilon_2/\rho_3) + (\exp(-\varepsilon_3/\rho_2) + (\exp(-\varepsilon_4/\rho_1) + \exp(-\varepsilon_5/\rho_1))^{\rho_1/\rho_2})^{\rho_2/\rho_3}]^{\rho_3}\}. \quad (2.5)$$

Define,

$$H(\mu, \rho) = \exp(\mu_1) + \{\exp(\mu_2/\rho_3) + [\exp(\mu_3/\rho_2) + (\exp(\mu_4/\rho_1) + \exp(\mu_5/\rho_1))^{\rho_1/\rho_2}]^{\rho_2/\rho_3}\}^{\rho_3}, \quad (2.6)$$

$$H_a(\mu, \rho) = \{\exp(\mu_2/\rho_3) + [\exp(\mu_3/\rho_2) + (\exp(\mu_4/\rho_1) + \exp(\mu_5/\rho_1))^{\rho_1/\rho_2}]^{\rho_2/\rho_3}\}^{\rho_3-1}, \quad (2.7)$$

$$H_b(\mu, \rho) = [\exp(\mu_3/\rho_2) + (\exp(\mu_4/\rho_1) + \exp(\mu_5/\rho_1))^{\rho_1/\rho_2}]^{\rho_2/\rho_3-1}, \quad (2.7)$$

$$H_c(\mu, \rho) = (\exp(\mu_4/\rho_1) + \exp(\mu_5/\rho_1))^{\rho_1/\rho_2-1}. \quad (2.8)$$

The probabilities of choosing alternative 1, 2, 3, 4, 5 can be derived as:

$$P_1(\mu, \rho) = \frac{\exp(\mu_1)}{H(\mu, \rho)}, \quad P_2(\mu, \rho) = \frac{H_a(\mu, \rho) \exp(\mu_2/\rho_3)}{H(\mu, \rho)}, \quad (2.9)$$

$$P_3(\mu, \rho) = \frac{H_a(\mu, \rho) H_b(\mu, \rho) \exp(\mu_3/\rho_2)}{H(\mu, \rho)}, \quad (2.10)$$

$$P_4(\mu, \rho) = \frac{H_a(\mu, \rho) H_b(\mu, \rho) H_c(\mu, \rho) \exp(\mu_4/\rho_1)}{H(\mu, \rho)}, \quad (2.11)$$

$$P_5(\mu, \rho) = \frac{H_a(\mu, \rho) H_b(\mu, \rho) H_c(\mu, \rho) \exp(\mu_5/\rho_1)}{H(\mu, \rho)}. \quad (2.12)$$

The log-likelihood function is,

$$\begin{aligned} L(\beta, \rho) = & \sum_{i=1}^N y_{i1} \log(P_1(\mu, \rho)) + \sum_{i=1}^N y_{i2} \log(P_2(\mu, \rho)) + \sum_{i=1}^N y_{i3} \log(P_3(\mu, \rho)) \\ & + \sum_{i=1}^N y_{i4} \log(P_4(\mu, \rho)) + \sum_{i=1}^N y_{i5} \log(P_5(\mu, \rho)), \end{aligned}$$

where y_{ij} ($i = 1, 2, \dots, N, j = 1, 2, \dots, M$) equals 1 if individual i chooses choice j , 0 otherwise. We assume that each equation contains K variables representing individual characteristics, i.e., X is a K dimensional vector with one variable as the constant, and β is a K dimensional vector of unknown coefficients. After one equation is normalized, there are $4K + 2$ coefficients to be estimated.

Although the theoretical framework of the NMNL model is not very complex, the actual maximization of the NMNL likelihood function has proved to be treacherous. Before Hensher (1986), it was widely recommended that sequential estimation (LIML) first be used, see Maddala (1983). Then, these LIML estimates were used as starting values for one-iteration of Newton-Raphson FIML step. The reasons for the two steps is that LIML estimates at higher levels are inefficient and their standard errors inconsistent due to the use of “estimates of estimates” in the estimation of the coefficients of inclusive values. The first iteration of FIML could correct this problem. Berkovec and Rust (1985) and Hensher (1986), however, reported that the log-likelihood and the parameter estimates tend to be very unstable after one iteration. Thus, it seems unwise to limit FIML to one iteration when the LIML estimates are used as starting values. On comparing LIML and FIML, many researchers including McFadden (1981) have observed that, even

in large samples, LIML parameters and their standard errors differ considerably from the FIML results. Kling and Thomson (1996) have demonstrated that welfare estimates resulting from NMNL models are very sensitive to the differences between LIML and FIML estimates.

Brownstone and Small (1989) reported results from a systematic comparison of LIML with FIML. They estimated a 2-level NMNL model using data on time-of-day choice for work trips. Their conclusion was: LIML is not efficient, linearized maximum likelihood estimation is better, and FIML is most preferable. However, their tree structure was very specific: “We tried more complex tree structure than the one described here, but they proved unstable and yielded impermissible values of the ρ parameters”, see Brownstone and Small (1989, p.68). Even with a simple tree structure, an ad hoc restriction that the inclusive value parameters are equal was imposed. As pointed out by Poirier (1996) the likelihood function of NMNL model is so ill-behaved that reliable estimation is not guaranteed by current ML algorithms.

3. Bayesian Analysis of the Nested Logit Model

Bayesian approach is natural for the NMNL models, given the difficulties and the common use of non-data-based information in ML estimation. Moreover, the use of reasonable informative priors may reduce the unreliability of the ML estimator. Poirier (1996) first studied a two-level NMNL model in a Bayesian framework, but did not suggest an algorithm to explore the posterior distribution. In this paper, we extend his work to higher level NMNL models, and develop a practical simulation framework for Bayesian inference.

3.1. Priors for the Parameters in the Indirect Utility Function

Poirier (1996) extended the Bayesian analysis of the multinomial logit (MNL) model (Koop and Poirier, 1993) to a NMNL model, where the distribution of errors is:

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_J) = \exp\{-\exp(-\varepsilon_m) - [\sum_{j \in J-m} \exp(-\varepsilon_j/\rho)]^\rho\}, \quad 0 < \rho < 1. \quad (3.1)$$

Assuming that ρ and β are independent, Poirier (1996) proposed the same priors for the parameters in the indirect utility function as those used in Koop and

Poirier (1993). Following Koop and Poirier's notation, the likelihood function of the MNL model is written as:

$$L(\beta) = f(\beta|G, R, Z) = \prod_{i=1}^N \prod_{r=1}^{R_i} \prod_{j=1}^{J_i} p_{ij}^{y_{irj}} = \frac{\exp(G'\beta)}{\prod_{i=1}^N \left[\sum_{j=1}^{J_i} \exp(z'_{ij}\beta) \right]^{R_i}}, \quad (3.2)$$

where $R \equiv [R_1, R_2, \dots, R_N]'$, R_i is the replications for i th individual and introduced here for controlling the strength of the priors. $G = \sum_{i=1}^N R_i \sum_{j=1}^{J_i} \bar{y}_{ij} z_{ij}$. $y_{irj} = 1$ iff i th individual choose j th alternative on r th replication, otherwise $y_{irj} = 0$. $\bar{y}_{ij} = R_i^{-1} \sum_{r=1}^{R_i} y_{irj}$, $Z = [z_1, z_2, \dots, z_i]'$, $z_i \equiv [z'_{i1}, z'_{i2}, \dots, z'_{iJ_i}]'$, z_{ij} contains the characteristics of both individual i and alternative j . The prior is specified as:

$$g(\beta) = \frac{f(\beta|\underline{G}, \underline{R}, Z)}{c(\underline{G}, \underline{R}, Z)}, \quad (3.3)$$

where $c(\underline{G}, \underline{R}, Z)$ is the normalization constant, $\underline{R} = [R_1, R_2, \dots, R_N]'$, and $\underline{G} = [\underline{G}'_1, \underline{G}'_2, \dots, \underline{G}'_J]'$ are prior hyperparameters. As discussed in Koop and Poirier (1993) and Poirier (1996), a 'neutral' case for the prior is to set $\underline{G} = 0_k$, which means equal probabilities for all alternatives when evaluated at the prior mode with $\rho = 1$. Further, the strength of the priors can be easily controlled by setting different values to \underline{R}_n in \underline{R} . Large values of \underline{R}_n means strong prior; as the values approach zero, the prior becomes a uniform distribution.

3.2. Priors for the Parameters of Inclusive Values

McFadden (1977, 1981) extended MNL to NMNL model by allowing the inclusive value (or dissimilarity) coefficient ρ to be different from one. It has been shown that ρ must lie in the unit interval $0 < \rho \leq 1$ for the empirical model to be globally consistent with utility maximization (Daly and Zachary 1979, and McFadden 1977). However, estimates of ρ greater than 1 are often encountered in practice. More recent work by Börsch-Supan (1990) and Koning and Ridder (1994) have derived local sufficiency conditions that permit values of $\rho > 1$ that are consistent with utility theory.

3.2.1. Poirier's Prior

Poirier (1996) proposed a class of priors for ρ , called the generalized logistic density:

$$f(\rho) = f_\rho(\rho|\underline{a}, \underline{b}, \underline{c}) = \frac{\underline{a}}{\underline{c}} \exp\left(\frac{\underline{b} - \rho}{\underline{c}}\right) \left[1 + \exp\left(\frac{\underline{b} - \rho}{\underline{c}}\right)\right]^{-(1+\underline{a})}. \quad (3.4)$$

The corresponding cumulative distribution function is:

$$F(\rho) = F_\rho(\rho|\underline{a}, \underline{b}, \underline{c}) = [1 + \exp(\frac{\underline{b} - \rho}{\underline{c}})]^{-\underline{a}}, \quad (3.5)$$

where \underline{a} , \underline{b} , \underline{c} are three hyperparameters. From this distribution, it is easy to verify that $E(\rho) = \underline{c}\{\ln[\Gamma(\underline{a})] - \ln[\Gamma(1)]\} + \underline{b}$, $Var(\rho) = \underline{c}^2\{\ln[\Gamma(\underline{a})]' - \ln[\Gamma(1)]'\}$, and the prior mode of ρ is $\underline{b} + \underline{c}\ln(\underline{a})$. To reduce the number of hyperparameters, Poirier (1996) assumed the prior probability under $H_1: \rho \neq 1$ concentrates in the neighborhood of $H_0: \rho = 1$ (MNL), and hence put the restriction $\underline{b} + \underline{c}\ln(\underline{a}) = 1$. In conjunction with other two conditions $P(0 < \rho < 1)$ and $P(\rho \geq 1)$, the three hyperparameters can be determined.

Although we follow Poirier (1996) for priors of the parameters in the deterministic utility function, we will suggest alternative priors for ρ . First, estimating the NMNL model means we assume that the data should be fitted to a nested logit, not multinomial logit model. The constraint that the prior probability under $H: \rho \neq 1$ concentrates in the neighborhood of $H_0: \rho = 1$ implies that the model should be MNL, not NMNL. Secondly, as discussed by Poirier (1996), his prior can not rule out $\rho \leq 0$.

3.2.2. Semi-flat Priors

It is simple to show that $\rho < 0$ implies that consumers are choosing the alternative to minimize their utilities³. Furthermore, the model cannot be defined when $\rho = 0$ because $\lim_{\rho \rightarrow -0} P_j \neq \lim_{\rho \rightarrow +0} P_j$ (P_j is the probability of choice j ; see appendix 1 for proof). Therefore, it is reasonable to restrict the range of the prior distribution of ρ to some interval such that the NMNL model is consistent with utility maximization. We restrict ρ to $(0, +\infty)$, because (1) the NMNL model can still be used as a valid statistical model when $\rho > 1$ (Train, Ben-Akiva and Atherton 1989; Amemiya and Shimono 1989), and 2) in applied work values of $\rho > 1$ have been very common. For instance, Hausman, Leonard and McFadden (1995) report FIML estimation of a 2-level recreational trip-allocation model where about half of over thirty inclusive value coefficients exceeded one by more than twice their standard errors.

We assume that ρ falls onto $(0, 1)$ with a probability λ , and given $0 < \rho < 1$, ρ has a flat distribution. We also assume that the probability of ρ being less than

³If $\rho < 0$, the Börsch-Supan (1990) procedure is not usable, because the density is globally negative when $-1 < \rho < 0$. See also Amemiya and Kim (1992),

0 is ξ and the probability of ρ far below 0 decreases exponentially, and we make the same assumption for $\rho > 1$. Thus we have the following prior density for ρ ,

$$f(\rho) = \begin{cases} \lambda \exp(\frac{\lambda}{\xi}\rho), & \text{if } \rho \leq 0, \\ \lambda, & \text{if } 0 < \rho < 1, \\ \lambda \exp[\frac{\lambda}{1-\lambda-\xi}(1-\rho)] & \text{if } \rho \geq 1. \end{cases} \quad (3.6)$$

We call this prior a “semi-flat” prior. A natural choice is to set $\lambda = 0.5$, which means that the likelihood of ρ falling onto the region $(0, 1)$ is 50%, and outside that region is also 50%. A special case is that the density function is also symmetric. This means that ρ has the same probability of being greater than 1 and less than 0,

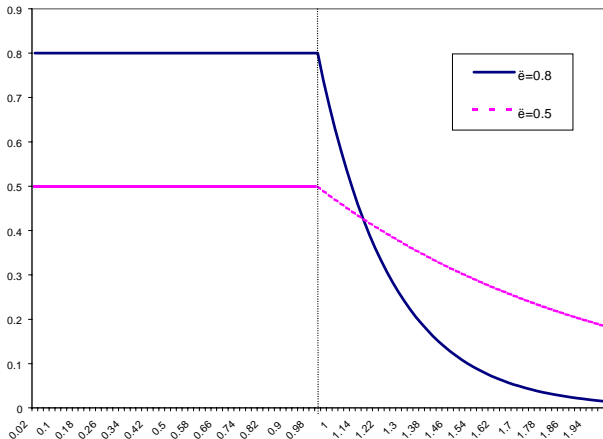
$$f(\rho) = \begin{cases} \frac{1}{2} \exp(2\rho), & \text{if } \rho \leq 0, \\ \frac{1}{2}, & \text{if } 0 < \rho < 1, \\ \frac{1}{2} \exp[2(1-\rho)] & \text{if } \rho \geq 1. \end{cases} \quad (3.7)$$

Even though the prior in (3.7) is not appropriate for NMNL models, the advantage of prior in (3.6) is that it allows for $P(\rho \leq 0) = 0$. For instance, set $\lambda = 0.5$, and $\xi = 0$, we have,

$$f(\rho) = \begin{cases} 0, & \text{if } \rho \leq 0, \\ \frac{1}{2}, & \text{if } 0 < \rho < 1, \\ \frac{1}{2} \exp(1-\rho) & \text{if } \rho \geq 1. \end{cases} \quad (3.8)$$

Figure 2 plots the semi-flat priors with $\xi = 0$, $\lambda = 0.5$ and $\lambda = 0.8$.

Figure 2. Semi-flat Priors



3.2.3. Sims' Priors

Sims (1991) suggested a prior in the context of unit root tests in time series analysis, see also Geweke (1994). This prior can also be used for ρ in the NMNL model. A generalization of Sims' prior is,

$$f(\rho) = \begin{cases} 0, & \text{if } \rho \leq 0, \\ \alpha(s)\rho^{s-a} \exp(-\rho^s), & \text{if } \rho > 0, \end{cases} \quad (3.9)$$

where s and a ($s \geq a$) are hyperparameters, and $\alpha(s)$ is the normalization constant. The prior mode of ρ is,

$$\rho = \left(\frac{s-a}{s} \right)^{\frac{1}{s}}. \quad (3.10)$$

It is easy to see that a and s control the prior location and variance. If $a > 0$, the prior mode of ρ is less than 1; if $a < 0$, the prior mode of ρ is greater than 1, and if $a = 0$, the prior mode of ρ equals 1, which means we believe that the data should be fitted to a MNL model. Setting $a = 1$, we have Sims' prior,

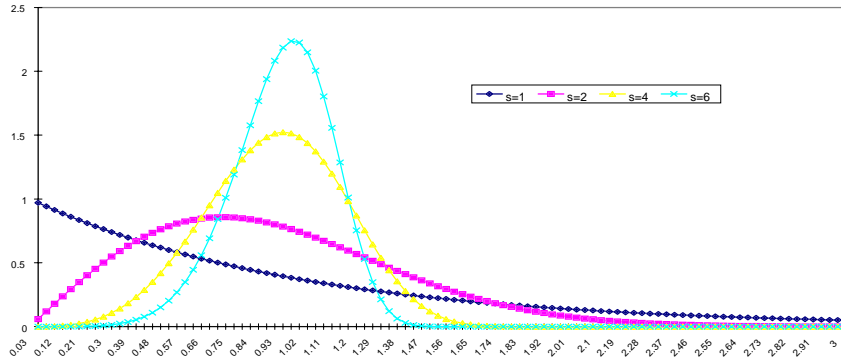
$$f(\rho) = \begin{cases} 0, & \text{if } \rho \leq 0, \\ s\rho^{s-1} \exp(-\rho^s), & \text{if } \rho > 0. \end{cases} \quad (3.11)$$

The corresponding cumulative distribution function is,

$$F(\rho) = \begin{cases} 0, & \text{if } \rho \leq 0, \\ 1 - \exp(-\rho^s). & \text{if } \rho > 0. \end{cases} \quad (3.12)$$

We plotted the density functions in (3.11) with $s = 1, 2, 4, 6$ in Figure 3⁴.

Figure 3. Sims' Prior



⁴In the case where the inclusive value parameter exceeds one, the procedure developed by Börsch-Supan (1990) was further extended by Kling and Herriges (1995) and Herriges and Kling (1996) to derive the exact upper limits C on ρ for a two-level NMNL model. The NMNL model remains compatible with utility-maximization as long as ρ is above zero but below the upper limit. Their calculations for a specific 2-level model suggest that the Bosch-Supan conditions will rarely be met for $\rho > 2$ and may not be met in many cases when $\rho < 2$. In the context of our priors this would suggest choosing appropriate values of hyperparameters λ (for the semi-flat prior) and s (for the Sims' prior). We do not recommend truncating the prior distribution of ρ using the calculated value of C since these estimates would embody the errors of model estimation and specification.

In addition to the priors discussed above, *Gamma* and *Beta* probability densities are also very attractive priors for ρ . For the NMNL model, noninformative priors are not ideal for ρ . Flat priors on different versions of the parameter space can yield different posterior distributions. For parameters having range from 0 to $+\infty$, Jeffreys suggested taking its logarithm to be uniform, $\theta = \ln(\rho)$, which is invariant to transformations of the form $\phi = \rho^n$, that is $d\phi = n\rho^{n-1}d\rho$, and thus $\frac{d\phi}{\phi} = \frac{d\rho}{\rho}$ (see Zellner, 1971, pp.41-53 for details).

3.3. The Posterior

For the MNL model, $g(\beta)$ in (3.3) is a natural conjugate prior, thus the posterior is simple:

$$g(\beta|y) = \frac{f(\beta|\bar{G}, \bar{R}, Z)}{c(\bar{G}, \bar{R}, Z)}, \quad (3.13)$$

where $\bar{G} = \underline{G} + G$, $\bar{R} = \underline{R} + R$. In the case of NMNL model, natural conjugate priors do not exist because there is no sufficient statistic for β (see Poirier, 1996). The posterior for the NMNL model is

$$\pi(\beta, \rho) = g(\beta)f(\rho)L(\beta, \rho) / \int g(\beta)f(\rho)L(\beta, \rho) d\beta d\rho,$$

where $f(\rho) = f(\rho_1)f(\rho_2)f(\rho_3)$ for a four-level model. For the parameters β in the utility function, as discussed in Koop and Poirier (1993), a ‘neutral’ case for the prior is to set $\underline{G} = 0_k$, which means equal probabilities for all alternatives when evaluated at the prior mode with $\rho = 1$. Further, the strength of the priors can be easily controlled by setting different values of \underline{R}_n in \underline{R} .

4. Metropolis-Hastings Algorithm and the Implementation Issues

4.1. The Metropolis-Hastings Algorithm in Logarithm

In order to understand the M-H algorithm, and its application to the NMNL posterior, the general state-space Markov chain theory is essential, especially the theory related to irreducibility, recurrence, aperiodicity, Harris recurrence, and ergodicity. We found Nummelin (1984) and Meyn and Tweedie (1993) to be excellent readings on this subject. For the purpose of this study, we only summarize

the basic concepts of M-H algorithm which are essential for the purpose of this paper.

For a transition kernel $P(x, A)$ of Markov chain $\{\Phi_n\}$ to converge to a stationary distribution ($\|P^n(x, \cdot) - \pi(\cdot)\| \rightarrow 0$)⁵ for π -almost all⁶ x , the chain must (1) be π -Irreducible, (2) be aperiodic, and (3) have a proper invariant distribution π , which satisfies, $\pi(A) = \int \pi(dx) P(x, A)$.

The irreducibility and aperiodic conditions for the Gibbs sampler and M-H algorithm are pure Markov chain issues. These conditions are usually satisfied if the proposal (or candidate-generating) distribution has a positive density on the same support as that of the target distribution. It is also satisfied by the proposal distribution with restricted support (see Chib and Greenberg, 1995)⁷. For our application of M-H algorithm (in which we employed Gaussian proposal distribution) to nested logit models, it is not difficult to verify that these conditions are satisfied.

The condition 3 is ensured by the introduction of a probability:

$$\alpha(x, y) = \begin{cases} \min \left\{ \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1 \right\}, & \text{if } \pi(x)q(x, y) > 0, \\ 1 & \text{if } \pi(x)q(x, y) = 0, \end{cases} \quad (4.1)$$

which is the heart of the M-H algorithm and is the key contribution of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), see also Hastings (1970). With the probability $\alpha(x, y)$, the transitional kernel

$$P(x, dy) = \alpha(x, y)q(x, y) dy + r(x) \delta_x(dy) \quad (4.2)$$

will converge to its invariant distribution π which is our target distribution⁸, and where $r(x) = 1 - \int p(x, y) dy$, δ_x denotes point mass at x , and $q(x, y)$ is the proposal distribution (candidate-generating density).

For very irregular posterior surfaces, it is helpful to rewrite the algorithm in the logarithm. The algorithm in logarithm can greatly reduce the numerical problems thereby avoid disrupting the simulation; it can also facilitate the computer

⁵ $P^n(x, A) = P\{x_n \in A | x_0 = x\}$, $x \in E$, $A \in E$, E is the state space, which often is K -dimensional Euclidean space IR^K , or can be more general.

⁶In fact, for M-H algorithm it is true for all x , since M-H kernels are Harris recurrent, see Tierney (1994).

⁷Also see Nummelin (1994), Roberts and Smith (1994), Tierney (1994) and Smith and Roberts (1993) for proofs.

⁸It needs to be pointed out that it is not necessary to have aperiodicity. What we are really concerned in practice is the convergence of sample path averages, which is guaranteed by condition 1 and 3.

program coding. The algorithm in logarithm can be written as:

$$\alpha^*(x, y) = \ln(\alpha(x, y)) = \min\{0, \ln(\pi(y)) - \ln(\pi(x)) + \ln(q(y|x)) - \ln(q(x|y))\}. \quad (4.3)$$

With a random number U from Uniform $(0, 1)$, if

$$\ln(U) \leq \alpha^*(x, y), \quad (4.4)$$

the candidate is accepted and one iteration is finished; if $\ln(U) > \alpha^*(x, y)$, the candidate is discarded and the chain does not move.

4.2. Implementation Issues

For complex posteriors, mixing speed of a chain is always crucial to MCMC practitioners, and improving the mixing efficiency needs endless effort. In this subsection we explain how the algorithm is tailored to ensure convergence of the chain.

4.2.1. Selection of the Chain

We adopted the random walk chain in this study. The random walk chain with proposal distribution $q(y|x) = q(|y - x|)$ is a special case of Metropolis chain with proposal distributions $q(y|x) = q(x|y)$ for all x and y . In this case,

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}, \quad (4.5)$$

which was proposed by Metropolis et al. (1953).

In fact, before using general M-H algorithm, we tried Block-at-a-Time Algorithms (see Chib and Greenberg, 1996; Geweke, 1999). Although it seems convenient to update each block of the parameters in different utility functions, ρ_1 and ρ_2 in turn, the Block-at-a-Time Algorithms did not deliver any advantage over the M-H algorithm. The Block-at-a-Time Algorithm in our context took 30% more time than the M-H algorithm.

We also tried independence chain with the proposal distribution $q(y|x) = q(y)$, which is a special case of M-H algorithm. It is well known that independence chain works well for posteriors with smooth surface, but does poorly for irregular posterior surfaces. In our own experience, the independence chain works well for the MNL model, but not for the NMNL model.

4.2.2. Reparameterization

Most of the algorithms are sensitive to the choice of parameterization of the model. For instance, the adaptive quadrature scheme by Naylor and Smith (1982) and the Laplace algorithms by Tierney and Kadane (1986) make some explicit assumptions about the shape of the posterior surface; the efficiency and accuracy of these algorithms are very sensitive to the choice of parameterization. The most compelling reason for the reparameterization in MCMC applications is to reduce the correlations among the coefficients, as we know that high correlations among coefficients will dramatically reduce the mixing speed. The parameterization of a statistical model may seem arbitrary, but the choice has both theoretical and practical significance in our case (see Hills and Smith, 1992). For the NMNL model, the reparameterization is necessary not only for reducing correlation among parameters, but also for stabilizing the simulation. The computer program will halt when these coefficients are near zero because the coefficients of the inclusive values appear in the denominator of the power. After an intensive search (also see Poirier, 1996), for the 4-level NMNL model (2.5) we use the following reparameterization:

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2/\rho_3, \quad \alpha_3 = \beta_3/\rho_2, \quad \alpha_4 = \beta_4/\rho_1,$$

$$\alpha_5 = \beta_5/\rho_1, \quad \gamma_1 = \rho_1/\rho_2, \quad \gamma_2 = \rho_2/\rho_3, \quad \gamma_3 = \rho_3.$$

Assuming that there are K variables in each equation (all β s are K dimensional), the determinant of the Jacobian matrix turns out to be $\gamma_1^{2K} \gamma_2^{3K+1} \gamma_3^{4K+2}$ (see appendix 2 for the proof). The empirical examples in this study are special cases of this four-level NMNL model. With this reparameterization, all the coefficients of the inclusive values in the denominators of the powers are removed and the computer program performs well.

4.2.3. Heating the Target

For a complicated nonlinear likelihood function, the high probability areas of the posterior are very likely to be separated by regions of very low probability. In this case, it is very difficult for the chain to travel from one mode to the other. Heating the target will flatten the surface of the posterior and make the chain more easily travel among the modes. Jennison (1993) gave a very simple example to explain how slow mixing can occur and suggests heating the target by τ ,

$$\pi^{(\tau)}(\beta) = \frac{1}{K(\tau)} \pi(\beta)^\tau. \quad (4.6)$$

where $1/\tau$ is called temperature, and $K(\tau)$ is a normalizing constant.

It should be noticed that heating target technique does not always work, the “witch’s hat” distribution is a well known example (see Geyer and Thompson, 1995; Matthews, 1993). For the NMNL model, heating is necessary, but overheating could result in extreme inefficiency. Jennison (1993) adopted $\tau = 4$ for his simple distribution. We tried $\tau = 5, 6, 7, 8, 9, 10$ for the NMNL model, and did not obtain the desired accuracy. We found $\tau = 3$ can significantly increase the mixing speed and maintains the needed sampling accuracy at the same time. Sometimes, overheating is useful. Overheating with overdispersed starting points of the chains can help researchers to avoid false convergence (see Gelman and Rubin, 1992a, 1992b).

4.3. Convergence Issues

For complex posterior surfaces, such as that of the NMNL model, the convergence issue (how long the chain should run) has been always been critical.⁹ However, the published literature has not suggested a universal diagnostic technique that can be applied safely to all models. Each diagnostic method has its own advantages and disadvantages. Currently, there are two schools of convergence diagnostic techniques, one is based on multiple chains (Gelman and Rubin, 1992a), and the other is based on a single long chain (Ripley, 1987; Geweke, 1992; Raftery and Lewis, 1992). Currently, the multiple-chains methods by Gelman and Rubin (1992a) and the single-chain methods by Raftery and Lewis (1992) are the most-often used techniques (Cowles and Carlin, 1996; Brooks and Roberts, 1999). The single-chain methods have been criticized for declaring false convergence (Gelman and Rubin, 1992b). The multiple-chains methods have been criticized for inefficiency since the “burn in”s have to be discarded.

In this study, we employ a 3-step procedure to examine the convergence of chains. This procedure combines the strengths of the multiple-chains and single-chain methods to ensure the convergence. In this procedure, the first and third steps are based on multiple-chains method, and the second step is based on single-chain method.

The steps 1 and 3 are based on Gelman and Rubin (1992a) and Cowles and Carlin (1996). Gelman and Rubin (1992a) technique (applied to $-2*\log(\text{posterior})$) is a reliable diagnostic method since it allows the chain start from different places.

⁹See Ripley (1987, Ch. 6), Geweke(1992, 1999), Raftery and Lewis (1992), Gelman and Rubin (1992a), Zellner and Min (1995), and Cowles and Carlin (1996).

As long as the starting distribution is well overdispersed, a dangerous situation that the chain is trapped in a neighborhood of one mode can be detected. This method monitors a “shrink factor” $\sqrt{\hat{R}}$ which declines to 1 as $n \rightarrow \infty$.

$$\sqrt{\hat{R}} = \sqrt{\frac{\nu \hat{V}}{\nu - 2W}}, \quad (4.7)$$

where $\hat{V} = \frac{n-1}{n}W + \frac{(m+1)}{mn}B$, $W = \frac{\sum S_i^2}{m}$, $S_i^2 = \frac{\sum (x_{ij} - \bar{x}_i)^2}{n-1}$, $B = \text{var}(\bar{x}_i)$, and $\nu = 2(\hat{V})^2 / \text{var}(\hat{V})$, and $\{x_{ij}\}$ are sequences of the scalar being monitored, with m being the number of sequences, each with length n .

In the next section, we present a number of illustrative NMNL models. Amongst them the labor supply Model I(1) of Amemiya and Shimono (1989) has the most complex posterior surface. We use this model as an example to illustrate how the three-step procedure is implemented.

Step 1: Run 40 chains independently ($m = 40$), with 25,000 iterations for each chain ($n = 25,000$). Assume that all the parameters are around 0's, and let chains start from $-2 + 0.1 * l$ ($l = 1, 2, \dots, 40$). This design enables the starting values of chains cover the ball $(-2, 2]^K$ (K is the number of dimensions of the posterior).

Step 2: Run 5 long chains ($N=1,400,000$) starting from widely different initial values to ensure there is no mode that the chain has not visited. There are compelling reasons for doing this. For the NMNL model, the likelihood function and, therefore, the posterior is highly nonlinear and non-concave (Amemiya and Kim, 1992; Poirier, 1996). Our concerns are (1) in step 1 the starting distribution may still not be overdispersed with respect to the posterior, (2) 40 chains may not be enough, and (3) therefore short chains (only 25,000 iterations) may omit one or more of the modes, e.g., before arriving at an isolated mode the chain has already reached 25,000 iterations. Moreover, in this step we also found “over heating” is helpful. Although the results reported in this study are based on $\tau = 3$, we did heat the posterior by $\tau = 5, 6, 7, 8, 9, 10$. Once τ is over 7, it can be seen that the chain is moving fast because the barriers among the modes are lifted. This ensures that the chains will not be stuck in the neighbor of one mode declaring false convergence. However, the over-heating technique can only be used to examine convergence. The over-heating is not an appropriate technique to produce final results because it reduces the algorithm's efficiency and can jeopardize the accuracy.

Step 3: Repeat step1, but we set the starting distribution as $-2 + \text{Mean} + 0.1 * l$,

where $Mean$ is the estimated mean of parameter vector of the posterior from step 1 or step 2. In this way, the chains cover the ball $(-2 + Mean, Mean + 2)^K$. This step plus step 1 and step 2 are designed to address the concern that the starting distribution may not be overdispersed.

For the models analyzed in the next section, the estimated shrink factors were all below 1.01. The current MCMC standard is that the shrink factor should be less than 1.20. Moreover, the results from the three steps were virtually the same.

5. Empirical Examples

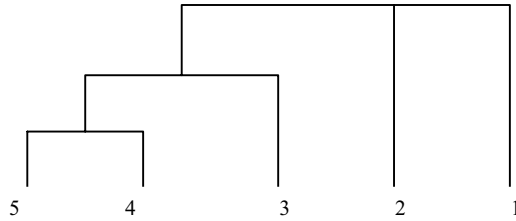
In this section we present two empirical examples to demonstrate how the M-H algorithm works for the NMNL model. The first example is adapted from Lahiri et al. (1995) describing the statutory structure of Social Security Administration (SSA)'s disability determination process, see also Hu et al. (2001). The second example is the model of labor supply behavior of the elderly from Amemiya and Shimono (1989).

5.1. Social Security's Disability Determination Process

SSA administers two disability programs providing cash assistance to persons who are unable to work due to health impairments: Disability Insurance (DI) and Supplemental Security Income (SSI). The determination process, which is used to screen applicants under both the programs, can be modeled parsimoniously using a three-level NMNL model. At the first level, the applicants are screened solely on the basis of the severity of the medical conditions. The most severely disabled applicants are accepted, the least severely disabled applicants are rejected and the indeterminate ones are passed on to the next level. At the second level, the remaining applications are evaluated in terms of their residual functional capacity and the demands of past occupation. If an applicant is judged to be able to do his/her past work despite the presence of impairments, the applicant will be rejected, and the remaining applicants will be passed on to the next stage for further evaluation. At the third level, the remaining applicants are judged on the basis of their ability to do any work, based on age, education, and work experience. If an applicant is judged to be able to do any work in the national economy, the applicant will be rejected, the rest of the candidates will be allowed

for disability benefits. The sequential structure, which is currently used across U.S. to evaluate nearly 2 million disability applications per year, is depicted in Figure 4:

Figure 4. Disability Determination Model



Thus, in this model, SSA adjudicators have five choices, and the applicants are sorted amongst these choices such that the following social welfare function is maximized (see McFadden 1975, 1976):

$$U_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad (5.1)$$

where $\mu_{ij} = X_i \beta_j'$, $j = 1, 2, \dots, 5$, $i = 1, 2, \dots, N$, N is the number of applicants. X_i is a K dimensional vector representing individual characteristics. Here choice 1 means denial for not having severe impairments, choice 2 means allowance due to meeting or equaling one of over 100 "listed" impairments, choice 3 means denial for being assessed to be able to do past work, choice 4 means allowance due to the inability to perform any work, and finally, choice 5 means denial due to the applicant's ability to do some work in the national economy.

As discussed in section 3, we adopted the prior specified in Poirier (1996) for the coefficients in the indirect utility functions, where the strength of the priors is controlled by \underline{R}_n . Koop and Poirier (1993, pp.332-338) presented a detailed sensitivity analysis for the MNL model. The results reported in this study are based on $\underline{R}_n = 0.2$. For the coefficients of the inclusive values, ρ_1 and ρ_2 , the

results reported here are based on the semi-flat prior and Sims' prior. For the semi-flat prior, ρ_1 and ρ_2 are assumed to be independent and have the following distribution ($l = 1, 2$):

$$f(\rho_l) = \begin{cases} 0, & \text{if } \rho_l \leq 0, \\ \lambda, & \text{if } 0 < \rho_l < 1, \\ \lambda \exp[\frac{\lambda}{1-\lambda}(1 - \rho_l)] & \text{if } \rho_l \geq 1. \end{cases} \quad (5.2)$$

In this study we use $\lambda = 0.5$.

For Sims' prior, we use $s = 2$:

$$f(\rho_l) = \begin{cases} 0, & \text{if } \rho_l \leq 0, \\ 2\rho_l \exp(-\rho_l^2), & \text{if } \rho_l > 0, \end{cases}$$

The simulation results are reported in Table 1¹⁰. The data set, based on an exact match between 1990 Survey of Income and Program Participation (SIPP) respondents and SSA's disability determination records, contains 1230 observations. We find that the parameters are estimated with precision and expected signs. The use of semi-flat compared to Sims' prior does not make much difference on these estimates. The two inclusive value parameters are small (approximately 0.03 and 0.14 respectively) and are estimated rather precisely. This means that, conditional on the covariates, the alternatives within each nest are very similar, but are dissimilar between groups. This evidence is consistent with the basic intent of the disability determination process.

5.2. The Labor Supply Behavior of the Elderly

Amemiya and Shimono (1989) extended a MNL model of labor supply behavior of the elderly to the NMNL model. They used seven independent variables in their study: age, amount of savings, amount of private pension, amount of public pension, other family members' income, dummy variable for health status (whether one considers oneself healthy or not), and a dummy variable representing whether an individual receives income other than wages and pension. The data set contains 4,101 observations.¹¹ For each individual, there are four choices: full-time

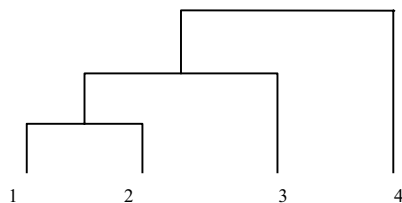
¹⁰Following Koop and Poirier (1993), we report the estimates of the posterior means and standard deviations. For convergence check, we used $m = 40$, $n = 20,000$, and $N = 1,200,000$. For all the long chains, 2% of total iterations were discarded as the initial transient stage (cf. Geyer 1992), and 20% of total iterations were discarded for the short chains (cf. Geweke 1999).

¹¹We are grateful to Professor Takeshi Amemiya for kindly sending the dataset to us.

employment (choice 1); part-time employment (choice 2); self-employment (choice 3); and retirement (choice 4). Amemiya and Shimono (1989) estimated 25 different models in four types. They considered Model Type I(1) to be the most plausible model, on which we implemented our MCMC. In order to compare how the number of nest levels affects the performance of the MCMC, we also analyzed Model III(1) which is a simple 2-level NMNL model.

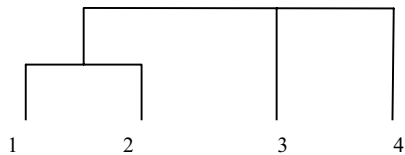
For type I(1) model, the tree structure is given in figure 5:

Figure 5. Labor Supply Behavior, Model I(1)



The tree structure for Model III(1) is depicted in figure 6:

Figure 6. Labor Supply Behavior, Model III(1)



We used the same priors as in the first example (i.e., the disability model). The simulation results for the model I(1) and III(1) are reported in Table 2 and Table 3 respectively. Here again the parameter estimates came up with signs consistent with expectations. The inclusive value parameters are found to be significantly greater than one, similar to those in Amemiya and Shimono (1989). However, these parameters are now estimated a lot more precisely.¹²

In order to demonstrate the unreliability of the ML estimation of the NMNL model, we also estimated the 2-level model III(1) using LIMDEP version 7 (Greene 1995, pp. 234-240.). The LIML and FIML estimates of ρ were 3.85 and 5.70 respectively. The standard error of ρ -estimate by MCMC was found to be considerably smaller than those from LIML/FIML estimations possibly due to the use of informative priors on ρ in the former application. The LIMDEP Manual reports estimation of a simple two-level NMNL model in which the FIML and LIML results are also quite different. We could never get LIMDEP or another popular software package TSP to estimate any NMNL model by FIML with more than two levels.

6. Conclusions

This study was motivated by the difficulties we encountered in estimating NMNL models by maximum likelihood methods. It is a common experience that FIML estimation of the NMNL model gives different results depending on the starting values, and sometimes the estimates never converge. Moreover, although both the LIML and FIML estimates are consistent, the LIML estimates are often substantially different from the FIML estimates, even in large samples.

In this paper we developed a practical MCMC algorithm for estimating the NMNL models in a Bayesian framework. We suggested two alternative informative priors on the inclusive value parameters that are fairly flexible to suit diverse empirical situations. Appropriate “heating target” and reparameterization techniques were adopted for fast mixing. We also found that the Metropolis-Hastings algorithm in logarithm can greatly reduce the numerical problems during the simulation. For illustrative purposes, we have implemented our algorithm on two real-life nested logit models involving fairly large datasets. The first example is Social Security’s 3-level disability determination process, Lahiri et al. (1995). The second one is taken from Amemiya and Shimono’s (1989) model of labor supply

¹²For model III(1), $m = 40$, $n = 15,000$, and $N = 900,000$ were enough to ensure convergence.

behavior of the aged. We applied a combination of various convergence criteria to ensure that the chain converged to its target distribution. Hopefully, our empirical examples demonstrate that the Bayesian approach, using Markov Chain Monte Carlo simulation techniques, is a viable alternative to analyze nested logit models.

References

- [1] Amemiya, T., 1985, *Advanced Econometrics*, (Cambridge, MA Harvard University Press).
- [2] Amemiya, T. and D. Kim, 1992, A Generalization of the Nested Logit Model, *Working paper*, Department of Economics, Stanford University.
- [3] Amemiya, T. and K. Shimono, 1989, An Application of Nested Logit Models to the Labor Supply of Elderly, *The Economic Studies Quarterly* 40, 14-22.
- [4] Berkovec, J. and J. Rust, 1985, A Nested Logit Model of Automobile Holdings for One-vehicle Households", *Transportation Reserach*, 19B, 275-285.
- [5] Besage, J. and P. J. Green, 1993, Spatial Statistics and Bayesian Computation, *Journal of the Royal Statistical Society, series B*, 55, 25-37.
- [6] Borsch-Supan, A. 1993, On the Compatibility of Nested Logit Models with Utility Maximization, *Journal of Econometrics* 43, 373-388.
- [7] Brooks, S. P. and G. O. Roberts, 1999, Assessing Convergence of Markov Chain Monte Carlo Algorithm, *Statistics and Computing*, forthcoming.
- [8] Brownstone, D. and K. A. Small, 1989, Efficient Estimation of Nested Logit Models, *American Statistical Association* 7, 67-74.
- [9] Cameron, T. A., 1985, A Nested Logit Model of Energy Conservation Activity by Owners of Existing Single Family Dwellings." *Review of Economics and Statistics*, 67, 205-211.
- [10] Casella, G. and E. I. George, 1992, Explaining the Gibbs Sampler, *The American Statistician* 46, 167-174.
- [11] Chib, S. and E. Greenberg, 1995, Understanding the Metropolis-Hastings Algorithm, *The American Statistician* 49, 327-335.

- [12] Chib, S. and E. Greenberg, 1996, Markov Chain Monte Carlo Simulation Methods in Econometrics, *Econometric Theory* 12, 409-431.
- [13] Chib, S. and E. Greenberg, 1996, Bayesian Analysis of Multivariate Probit Models, *Working Paper* (John M. Olin school of Business, Washington University).
- [14] Chib, S., E. Greenberg and Y. Chen, 1998, MCMC Methods for Fitting and Comparing Multinomial Response Models, *Working Paper* (John M. Olin school of Business, Washington University).
- [15] Cowles, M. K. and B. P. Carlin, 1996, Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association* 91, 883-904.
- [16] Daly, A. and S. Zachary, 1979, Improved Multiple Choice Models, in: David Hensher and Q. Dalvi, eds., *Identifying and Measuring the Determinants of Model Choice*, pp.335-357 (Teakfield, London).
- [17] Dubin, J.A., 1998, *Studies in Consumer Demand—Econometric Methods Applied to Market Data*, Kluwer Academic Publishers.
- [18] Falaris, E. M., 1984, A Model of Occupational Choice, *Research in Population Economics* 5, 289-307.
- [19] Falaris, E. M., 1987, A Nested Logit Migration Model with Selectivity, *International Economic Review* 28, 429-443.
- [20] Gelfand, E. A., S. K. Sahu and B. P. Carlin, 1995, Efficient Parameterization for Normal Linear Mixed Models, *Biometrika* 82, 479-488.
- [21] Gelman, A., and D. B. Rubin, 1992a, Inference from Iterative Simulation using Multiple Sequences (with discussions), *Statistical Science*, 7, 457-511.
- [22] Gelman, A., and D. B. Rubin, 1992b, A Single Sequence from the Gibbs Sampler Gives a False Sense of Security, in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 625-631.
- [23] Geweke, J., 1989, Bayesian Inference in Econometric Models Using Monte Carlo Integration, *Econometrica* 57, 1317-1339.

- [24] Geweke, J., 1992, Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Simith), 169-193.
- [25] Geweke, J., 1994, Priors for Macroeconomic Time Series and Their application, *Econometric Theory*, 10, 609-632.
- [26] Geweke, J., 1999, Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication, *Econometric Reviews* 18, 1-126.
- [27] Geyer, C. J., 1992, Practical Markov Chain Monte Carlo, *Statistical Science* 7, 473-511.
- [28] Geyer, C. J., and E. A. Thompson ,1995, Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference, *Journal of the American Statistical Association* 90, No.431, 909-920.
- [29] Greene, W., 1995, LIMDEP, Version 7.0: User's manual. Bellport, N.Y.: Econometric Software.
- [30] Guimaraes, P., R.J. Robert, and W.P. Douglas, 1998, Regional Incentives and Industrial Location in Puerto Rico, *International Regional Science Review* 21, 119-138.
- [31] Hastings, W. K., 1970, Monte Carlo Sampling Methods using Markov Chains And Their Applications, *Biometrika* 57, 97-109.
- [32] Hausman, J.A., G.K. Leonard, and D. McFadden, 1995, A Utility-Consistent Discrete Choice and Count Data Model: Assessing Recreational Use Losses Due to National Resource Damage, *Journal of Public Economics*, 56, 1-30.
- [33] Hensher, D. A., 1986, Sequential and Full Information Maximum Likelihood Estimation of a Nested Logit Model, *The Review of Economics and Statistics*, 657-667.
- [34] Herriges, J.A. and Kling, C.L., 1996, Testing the Consistency of Nested Logit Models with Utility Maximization, *Economics Letters* 50, 33-39.
- [35] Hills, S. E. and A. F. M. Smith, 1992, Parameterization Issues in Bayesian Inference, *Bayesian Statistics* 4, 227-246.

- [36] Hoffman, S. D. and G. J. Duncan, 1988, A Comparison of Choice-Based Multinomial and Nested Logit Models, *The Journal of Human Resources*, 550-602.
- [37] Hu, J., K. Lahiri, D. Vaughan, and B. Wixon, 2001, A Structural Model of Social Security's Disability Determination Process, *Review of Economics and Statistics*, May, Forthcoming.
- [38] Jennison, C. 1993, Discussion on the Meeting on the Gibbs Sampler and other Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society, Series B*, 55, 54-56.
- [39] Kling, C.L. and C.J. Thomson, 1996, The Implications of Model Specification for Welfare Estimation in Nested Logit Models, *American Journal of Agricultural Economics* 78, 103-114.
- [40] Kling, C.L. and J.A., Herriges, 1995, An Empirical Investigation of the Consistency of Nested Logit Models with Utility Maximization, *American Journal of Agricultural Economics* 77, 875-884.
- [41] Koning, R.H. and Ridder, G., 1994, On the Compatibility of Nested Logit Models with Utility Maximization-A Comment, *Journal of Econometrics* 63, 389-396.
- [42] Koop, G. and D. J. Poirier, 1993, Bayesian Analysis of Logit Models using Natural Conjugate Priors, *Journal of Econometrics* 56, 323-340.
- [43] Lahiri, K., D. R. Vaughan and B. Vixon, 1995, Modeling SSA's Sequential Disability Determination Process using Matched SIPP Data, *Social Security Bulletin* 58, 3-42.
- [44] Maddala, G. S., 1983, *Limited Dependent and Qualitative Variables in econometrics*, Cambridge University Press.
- [45] Matthews, P., 1993, A Slowly Mixing Markov Chain with Implications for Gibbs Sampling, *Statistics and Probability Letters*, 17, 231-236.
- [46] McFadden, D., 1975, The Revealed Preference of a Government Bureaucracy: Theory, *The Bell Journal of Economics*, 6, 401-416.

- [47] McFadden, D., 1976, The Revealed Preference of a Bureaucracy: Empirical Evidence, *The Bell Journal of Economics*, 7, 55-72.
- [48] McFadden, D., 1977, Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments, *Cowles Foundation Discussion Paper*, No.474.
- [49] McFadden, D., 1978, Modelling the Choice of Residential Location, in: *Spatial Interaction Theory and Residential Location*, 75-96, eds., A. Karlqvist et al. (North Holland: Amsterdam).
- [50] McFadden, D., 1980, Econometric Models for Probabilistic Choice among Products, *Journal of Business* 53, S13-S29.
- [51] McFadden, D., 1981, Econometric Models of Probabilistic Choice, in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, pp.198-272, Cambridge Mass.: MIT Press.
- [52] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, (1953), Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1092.
- [53] Meyn, S. P. and R. L. Tweedie, 1993, *Markov Chains and Stochastic Stability* (Springer-Verlag London Limited).
- [54] Naylor, J. C. and A. F. M. Smith, 1982, Applications of a Method for the Efficient Computation of Posterior Distributions, *Annals of Statistics*, 31, 214-225.
- [55] Newbold, K.B., 1997, Primary, Return and Onward Migration in the U.S. and Canada: Is There a Difference? *Papers in Regional Science* 76, 175-198.
- [56] Nummelin, E., *General Irreducible Markov Chains and Nonnegative Operators*, 1984, Cambridge, Cambridge University Press.
- [57] Poirier, D. J., 1996, A Bayesian Analysis of Nested Logit Models, *Journal of Econometrics* 75, 163-181.
- [58] Raftery, A. E. and S. M. Lewis, 1992, How Many Iterations in the Gibbs Sampler?, in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Simith), 763-773.

- [59] Ripley, B. D., *Stochastic Simulation*, 1987, New York, Wiley.
- [60] Roberts, G. O. and A. F. M. Smith, 1994, Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms, *Stochastic Processes and Their Applications* 49, 207-216.
- [61] Roberts, G. O. and R. L. Tweedie, 1994, Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms, *Research Report No. 94-9*, University of Cambridge, Statistical Laboratory.
- [62] Sims C. A., 1991, Comment on "To criticize the Critics" by Peter C. B. Phillips, *Journal of Applied Econometrics*, 423-434.
- [63] Smith, A. F. M., and G. O. Roberts, 1993, Bayesian Computation via The Gibbs Sampler and Related Markov Chain Monte Carlo Methods, *J. R. Statist. Soc. B* 55, 3-23.
- [64] Tierney, L., 1994, Markov Chain for Exploring Posterior Distributions, *The Annals of Statistics* 4, 1701-1762.
- [65] Tierney, L., and J. B. Kadane, 1986, Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86.
- [66] Train, K.E., 1980, A Structured Logit Model of Auto Ownership and Choice, *Review of Economic Studies*, 357-369.
- [67] Train, K.E., Ben-Akiva, M., and T. Atheron, 1989, Consumption Patterns and Self-Selecting Tariffs", *Review of Economics and Statistics*, 71, (January), 62-73.
- [68] Weiler, W. C., 1989, A Flexible Approach to Modeling Enrollment Choice Behavior, *Economics Of Education Review* 8, 277-283.
- [69] Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics*, (Wiley, New York, N.Y.), reprinted in 1987 (Krieger Publishing Co., Malabar, FL.).
- [70] Zellner, A. and C. Min, 1995, Gibbs Sampler Convergence Criteria, *Journal of The American Statistical Association* 90, 921-927.

Appendix 1:

Discontinuity of the NMNL likelihood function at $\rho = 0$:

Suppose a simple two-level NMNL model with three choices has the distribution of error terms as the following

$$f(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\{-\exp(-\varepsilon_3) - [\exp(-\varepsilon_2/\rho) + \exp(-\varepsilon_1/\rho)]^\rho\}. \quad (6.1)$$

The probabilities of the three choices are,

$$P_1(\mu) = \frac{\exp(\mu_1/\rho)(\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^{\rho-1}}{\exp(\mu_3) + (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho}, \quad (6.2)$$

$$P_2(\mu) = \frac{\exp(\mu_2/\rho)(\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^{\rho-1}}{\exp(\mu_3) + (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho}, \quad (6.3)$$

$$P_3(\mu) = \frac{\exp(\mu_3)}{\exp(\mu_3) + (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho}. \quad (6.4)$$

Here we only prove the corner solution when $\rho \rightarrow 0$ for the probability of choice 1,

$$\lim_{\rho \rightarrow +0} P_1(\mu) = \begin{cases} e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 > \mu_2, \\ \frac{1}{2}e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 = \mu_2, \\ 0, & \text{if } \mu_1 < \mu_2. \end{cases}$$

$$\lim_{\rho \rightarrow -0} P_1(\mu) = \begin{cases} 0, & \text{if } \mu_1 > \mu_2, \\ \frac{1}{2}e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 = \mu_2, \\ e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 < \mu_2. \end{cases}$$

$$\begin{aligned} P_1(\mu) &= \frac{\exp(\mu_1/\rho)(\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^{\rho-1}}{\exp(\mu_3) + (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho} \\ &= \frac{\frac{\exp(\mu_1/\rho)}{\exp(\mu_2/\rho) + \exp(\mu_1/\rho)}(\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho}{\exp(\mu_3) + (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho} = \frac{AB}{\exp(\mu_3) + B}. \end{aligned}$$

If $\mu_1 > \mu_2$,

$$\lim_{\rho \rightarrow +0} A = \lim_{\rho \rightarrow +0} \frac{\exp(\mu_1/\rho)}{\exp(\mu_2/\rho) + \exp(\mu_1/\rho)} = \lim_{\rho \rightarrow +0} \frac{1}{\exp((\mu_2 - \mu_1)/\rho) + 1} = 1.$$

$$\begin{aligned}\lim_{\rho \rightarrow +0} B &= \lim_{\rho \rightarrow +0} (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho = \exp(\mu_1) \lim_{\rho \rightarrow +0} (1 + \exp(\frac{\mu_2 - \mu_1}{\rho}))^\rho \\ &= \exp(\mu_1),\end{aligned}$$

$$\lim_{\rho \rightarrow +0} P_1(\mu) = \frac{\exp(\mu_1)}{\exp(\mu_1) + \exp(\mu_3)}.$$

$$\lim_{\rho \rightarrow -0} A = \lim_{\rho \rightarrow -0} \frac{\exp(\mu_1/\rho)}{\exp(\mu_2/\rho) + \exp(\mu_1/\rho)} = \lim_{\rho \rightarrow -0} \frac{1}{\exp((\mu_2 - \mu_1)/\rho) + 1} = 0.$$

$$\begin{aligned}\lim_{\rho \rightarrow -0} B &= \lim_{\rho \rightarrow -0} (\exp(\mu_2/\rho) + \exp(\mu_1/\rho))^\rho = \exp(\mu_2) \lim_{\rho \rightarrow -0} (1 + \exp(\frac{\mu_1 - \mu_2}{\rho}))^\rho \\ &= \exp(\mu_2),\end{aligned}$$

$$\lim_{\rho \rightarrow -0} P_1(\mu) = 0.$$

If $\mu_1 = \mu_2$,

$$\lim_{\rho \rightarrow 0} B = \exp(\mu_1),$$

$$\lim_{\rho \rightarrow 0} A = \frac{1}{2},$$

$$\lim_{\rho \rightarrow 0} P_1(\mu) = \frac{1}{2} \frac{\exp(\mu_1)}{\exp(\mu_1) + \exp(\mu_3)}.$$

If $\mu_2 > \mu_1$,

$$\lim_{\rho \rightarrow +0} A = 0, \quad \lim_{\rho \rightarrow +0} B = \exp(\mu_2),$$

$$\lim_{\rho \rightarrow +0} P_1(\mu) = 0. \quad \lim_{\rho \rightarrow -0} A = 1,$$

$$\lim_{\rho \rightarrow -0} B = \exp(\mu_1), \quad \lim_{\rho \rightarrow -0} P_1(\mu) = \frac{\exp(\mu_1)}{\exp(\mu_1) + \exp(\mu_3)}.$$

Therefore we have the results,

$$\lim_{\rho \rightarrow +0} P_1(\mu) = \begin{cases} e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 > \mu_2, \\ \frac{1}{2}e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 = \mu_2, \\ 0, & \text{if } \mu_1 < \mu_2. \end{cases}$$

$$\lim_{\rho \rightarrow -0} P_1(\mu) = \begin{cases} 0, & \text{if } \mu_1 > \mu_2, \\ \frac{1}{2}e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 = \mu_2, \\ e^{\mu_1}/\{e^{\mu_1} + e^{\mu_3}\}, & \text{if } \mu_1 < \mu_2. \end{cases}$$

Appendix 2:

Determinant of the Jakobian Matrix for the Reparameterization:

We set $\alpha_1 = \beta_1$, $\alpha_2 = \beta_2/\rho_3$, $\alpha_3 = \beta_3/\rho_2$, $\alpha_4 = \beta_4/\rho_1$, $\alpha_5 = \beta_5/\rho_1$, $\gamma_1 = \rho_1/\rho_2$, $\gamma_2 = \rho_2/\rho_3$, $\gamma_3 = \rho_3$, where α 's and β 's are K dimensional vectors, γ 's and ρ 's are scalars. Denote $\theta = (\beta, \rho)'$ and $\eta = (\alpha, \gamma)'$, which are $5K + 3$ dimensional. Define

$$J = \left| \frac{\partial \theta}{\partial \eta} \right| = |J_{l,m}|_{(5K+3) \times (5K+3)}.$$

$J_{l,m} = 0$, if $m > l$, or $m < l$ and $l \leq 5K$.

$J_{l,l} = \gamma_1^{I[l=3K+1, \dots, 5K]} \gamma_2^{I[l=2K+1, \dots, 5K]} \gamma_3^{I[l=K+1, \dots, 5K]} E$, if $l = m = 1, 2, \dots, 5K$,

$J_{l,l} = \gamma_2^{I[l=5K+1]} \gamma_3^{I[l=5K+1, 5K+2]}$, if $l = m = 5K + 1, 5K + 2, 5K + 3$.

$J_{5K+1,m} = I[m = 3K + 1, \dots, 5K](\alpha_m \gamma_2 \gamma_3)$, if $1 \leq m \leq 5K$.

$J_{5K+2,m} = I[m = 2K+1, \dots, 5K+1] \alpha_m^{I[m=2K+1, \dots, 5K]} \gamma_1^{I[m=3K+1, \dots, 5K+1]} \gamma_3^{I[m=2K+1, \dots, 5K+1]}$,
if $1 \leq m \leq 5K + 1$.

$J_{5K+3,m} = I[m = K+1, \dots, 5K+2] \alpha_m^{I[m=K+1, \dots, 5K]} \gamma_1^{I[m=3K+1, \dots, 5K+1]} \gamma_2^{I[m=2K+1, \dots, 5K+2]}$,
if $1 \leq m \leq 5K + 2$.

$I[\cdot]$ is the indicator function, E is a $K \times K$ identity matrix.

Evaluating the determinant J , we get, $|J| = \gamma_1^{2K} \gamma_2^{3K+1} \gamma_3^{4K+2}$.

Appendix 3:

Variable definitions.

SSA's Disability model (Lahiri et al. 1995):

Mentdisd—Any one of five mental disabilities or a mental condition reported as causing a work or activity limitation.

T9100w3d—At least 1 overnight hospital stay in the last 12 months.

Work90c—Recent work experience and disability determination occurred in 1990.

Sal36—Three or more severe ADLs, wave 6.

Sil13—one or more severe IADLs, wave 3.

Sexd—Gender (male).

Noworkd—No recent work experience.

Workv2d2—Work limited, but able to perform prior work (both in Wave 2).

Occsipp3—Principal occupation of prior work was in sales or service.

Msf—Never Married.

Age56—Aged 55 or older (18-54 in base).

Yondment—Under age 35 and has a mental condition.

T8800w6b—General health status very good (wave 6).

T8800w6e—General health status poor (wave 6).

Young—Aged 18-34 (35 plus in base).

β_{jk} is the coefficient in j th equation for j th choice, $j = 1, 2, 3, 4$.

The equation corresponding choice 5 is normalized.

Labor supply model (Amemiya and Shimono, 1989):

Age—actual age, from 55 to 69.

Health —dummy for health status, equals 1 if healthy, otherwise 0.

Assets—amount of savings.

Pripen—amount of private pension.

Pubpen—amount of public pension.

Inothwp—dummy for an individual receiving income other than wages/pension.

Othfmin—other family members' income.

The equation corresponding choice 1 is normalized.

Table 1: Simulation Results for the Disability Determination Model

Variables	Coefficients	Semi-flat Prior		Sims' Prior	
		Mean	Std. dev.	Mean	Std. dev.
Choice1, denial: impairment not severe					
Constant	β_{10}	-1.211	0.112	-1.217	0.092
Mentdisd	β_{11}	-0.422	0.282	-0.296	0.190
T9100w3d	β_{12}	-0.220	0.185	-0.196	0.183
Work90c	β_{13}	0.456	0.283	0.560	0.211
Sal36	β_{14}	-0.877	0.948	-0.625	0.628
Sil13	β_{15}	-0.557	0.291	-0.536	0.231
Choice 2, allowance: meets or equals the listings					
Constant	β_{20}	-1.093	0.105	-1.077	0.085
Mentdisd	β_{21}	0.450	0.189	0.384	0.140
T9100w3d	β_{22}	0.294	0.166	0.315	0.145
Work90c	β_{23}	-0.605	0.284	-0.617	0.239
Sal36	β_{24}	0.685	0.279	0.698	0.280
Sil13	β_{25}	0.350	0.146	0.313	0.150

Table 1 continued.

Variables	Coefficients	Mean	Std. dev.	Mean	Std. dev
Choice3, denial: can do past work					
Constant	β_{30}	-0.078	0.060	-0.087	0.037
Sexd	β_{31}	-0.062	0.040	-0.070	0.031
Noworkd	β_{32}	-0.060	0.065	-0.074	0.042
Workv2d2	β_{32}	0.089	0.068	0.116	0.054
Occsipp3	β_{34}	0.065	0.046	0.059	0.037
Msf	β_{35}	-0.157	0.081	-0.145	0.055
Choice 4, allowance: cannot do some work					
Constant	β_{40}	-0.029	0.014	-0.029	0.011
Age56	β_{41}	0.060	0.028	0.061	0.021
Yondment	β_{42}	0.077	0.038	0.075	0.028
T8800w6b	β_{43}	-0.012	0.015	-0.016	0.012
T8800w6e	β_{44}	0.012	0.009	0.012	0.008
Age12	β_{45}	-0.024	0.018	-0.027	0.015
Coefficients of Inclusive Values					
Inclusive value	ρ_1	0.028	0.014	0.029	0.010
Inclusive value	ρ_2	0.141	0.069	0.153	0.040

Note: See appendix for variable definitions.

Table 2: Simulation Results from Amemiya-Shimono (1989) Model I(1)

Variables	Coefficients	Semi-flat Prior		Sims' Prior	
		Mean	Std. dev.	Mean	Std. dev.
Choice 2: Part Time					
Constant	β_{20}	-5.951	0.626	-5.708	0.589
Age	β_{21}	1.392	0.354	1.381	0.342
Health	β_{22}	-0.800	0.271	-0.805	0.283
Pripen	β_{23}	0.542	0.264	0.543	0.240
Pubpen	β_{24}	2.504	0.404	2.401	0.345
Inothwp	β_{25}	0.551	0.240	0.488	0.242
Choice 3: Self-employed					
Constant	β_{30}	-6.721	0.711	-6.528	0.609
Age	β_{31}	2.987	0.427	2.904	0.379
Health	β_{32}	-0.188	0.289	-0.183	0.270
Assets	β_{33}	0.637	0.274	0.630	0.248
Pripen	β_{34}	0.390	0.380	0.410	0.365
Pubpen	β_{35}	0.744	0.341	0.751	0.309
Inothwp	β_{36}	0.461	0.374	0.449	0.376

Table 2 Continued

Variables	Coefficients	Semi-flat Prior		Sims' Prior	
		Mean	Std. dev.	Mean	Std. dev.
Choice 4: Retired					
Constant	β_{40}	-4.952	0.245	-4.923	0.262
Age	β_{41}	2.644	0.215	2.621	0.215
Health	β_{42}	-1.975	0.141	-1.981	0.140
Assets	β_{43}	-0.285	0.167	-0.265	0.159
Pripen	β_{44}	0.755	0.141	0.724	0.148
Pubpen	β_{45}	3.694	0.213	3.669	0.200
Inothwp	β_{46}	0.940	0.139	0.912	0.132
Othfmin	β_{47}	0.649	0.154	0.654	0.150
Inclusive Value	ρ_1	1.824	0.168	1.747	0.149
Inclusive Value	ρ_2	2.125	0.179	2.059	0.162

Note: See appendix 3 for variable definitions.

Table 3: Simulation Results from Amemiya (1989) Model III(1)

Variables	Coefficients	Semi-flat Prior		Sims' Prior	
		Mean	Std. dev.	Mean	Std. dev.
Choice 2: Part Time					
Constant	β_{20}	-5.572	0.551	-5.489	0.541
Age	β_{21}	1.355	0.320	1.337	0.327
Health	β_{22}	-0.750	0.256	-0.769	0.260
Pripen	β_{23}	0.570	0.243	0.512	0.221
Pubpen	β_{24}	2.351	0.312	2.300	0.319
Inothwp	β_{25}	0.523	0.254	0.541	0.249
Choice 3: Self-employed					
Constant	β_{30}	3.288	0.188	-3.275	0.167
Age	β_{31}	1.595	0.169	1.570	0.160
Health	β_{32}	-0.252	0.134	-0.250	0.142
Assets	β_{33}	0.324	0.138	0.319	0.133
Pripen	β_{34}	0.264	0.189	0.267	0.206
Pubpen	β_{35}	0.587	0.159	0.585	0.150
Inothwp	β_{36}	0.335	0.203	0.355	0.201

Table 3 Continued

Variables	Coefficients	Semi-flat Prior		Sims' Prior	
		Mean	Std. dev.	Mean	Std. dev.
Choice 4: Retired					
Constant	β_{40}	-4.941	0.249	-4.908	0.247
Age	β_{41}	2.495	0.190	2.485	0.198
Health	β_{42}	-1.950	0.139	-1.969	0.142
Assets	β_{43}	-0.300	0.144	-0.321	0.151
Pripen	β_{44}	0.719	0.136	0.726	0.143
Pubpen	β_{45}	3.655	0.209	3.639	0.207
Inothwp	β_{46}	0.893	0.131	0.924	0.136
Othfmin	β_{47}	0.662	0.161	0.670	0.155
Inclusive Value	ρ	1.677	0.140	1.649	0.141

Note: See appendix 3 for variable definitions.

Table 4: ML Results from Amemiya-Shimono (1989) Model III(1)

Variables	Coefficients	LIML		FIML	
		Estimate	Std. dev.	Estimate	Std. dev.
Choice 2: Part Time					
Constant	β_{20}	-4.368	0.215	-4.081	0.206
Age	β_{21}	1.018	0.196	1.082	0.181
Health	β_{22}	-0.552	0.123	-0.301	0.101
Pripen	β_{23}	0.353	0.141	0.423	0.109
Pubpen	β_{24}	1.633	0.174	1.648	0.165
Inothwp	β_{25}	0.491	0.152	0.336	0.123
Choice 3: Self-employed					
Constant	β_{30}	-4.143	0.208	-4.077	0.221
Age	β_{31}	1.935	0.179	2.121	0.229
Health	β_{32}	-0.607	0.120	-0.525	0.145
Assets	β_{33}	0.349	0.129	0.363	0.134
Pripen	β_{34}	0.326	0.132	0.455	0.180
Pubpen	β_{35}	1.275	0.231	1.637	0.297
Inothwp	β_{36}	0.537	0.145	0.512	0.181

Table 4 Continued

Variables	Coefficients	LIML		FILM	
		Estimate	Std. dev.	Estimate	Std. dev.
Choice 4: Retired					
Constant	β_{40}	-9.754	0.471	-9.857	0.546
Age	β_{41}	3.312	0.277	3.662	0.371
Health	β_{42}	-1.977	0.142	-1.854	0.170
Assets	β_{43}	-0.283	0.159	-0.262	0.150
Pripen	β_{44}	0.982	0.153	1.194	0.218
Pubpen	β_{45}	4.856	0.360	5.356	0.477
Inothwp	β_{46}	1.432	0.170	1.393	0.219
Othfmin	β_{47}	0.036	0.008	0.036	0.007
Inclusive Value	ρ	3.854	0.873	5.695	1.479

Note: See appendix 3 for variable definitions.